Felice Dell'Orletta, Johanna Monti and Fabio Tamburini (dir.)

Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020
Bologna, Italy, March 1-3, 2021

Accademia University Press

Datasets and Models for Authorship Attribution on Italian Personal Writings

Gaetana Ruggiero, Albert Gatt and Malvina Nissim

DOI: 10.4000/books.aaccademia.8880
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2020
Published on OpenEdition Books: 3 September 2021
Series: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic EAN: 9791280136336

Datasets and Models for Authorship Attribution on Italian Personal Writings

Gaetana Ruggiero[•◇], **Albert Gatt**[•], **Malvina Nissim**[◇]

[•]Institute of Linguistics and Language Technology, University of Malta, Malta
[◇]Center for Language and Cognition, University of Groningen, The Netherlands
garuggiero@gmail.com, albert.gatt@um.edu.mt, m.nissim@rug.nl

Abstract

Existing research on Authorship Attribution (AA) focuses on texts for which a lot of data is available (e.g novels), mainly in English. We approach AA via Authorship Verification on short Italian texts in two novel datasets, and analyze the interaction between genre, topic, gender and length. Results show that AV is feasible even with little data, but more evidence helps. Gender and topic can be indicative clues, and if not controlled for, they might overtake more specific aspects of personal style.

1 Introduction and Background

Authorship Attribution (AA) is the task of identifying authors by their writing style. In addition to being a tool for studying individual language choices, AA is useful for many real-life applications, such as plagiarism detection (Stamatatos and Koppel, 2011), multiple accounts detection (Tsikerdekis and Zeadally, 2014), and online security (Yang and Chow, 2014).

Most work on AA focuses on English, on relatively long texts such as novels and articles (Juola, 2015) where personal style could be mitigated due to editorial interventions. Furthermore, in many real-world applications the texts of disputed authorship tend to be short (Omar et al., 2019).

The PAN 2020 shared task was originally meant to investigate multilingual AV in fanfiction, focusing on Italian, Spanish, Dutch and English (Bevendorff et al., 2020). However, the datasets were eventually restricted to English only, to maximize the amount of available training data (Kestemont et al., 2020), emphasizing the difficulty in compiling large enough datasets for less-resourced languages.

AA research in Italian has largely focused on the single case of Elena Ferrante (Tuzzi and Cortelazzo, 2018) [1]. The present work seeks a more realistic take, using more diverse, user-generated data namely web forums comments and diary fragments, thereby introducing two novel datasets for this task: *ForumFree* and *Diaries*.

We cast the AA problem as *authorship verification* (AV). Rather than identifying the specific author of a text (the most common task in AA), AV aims at determining whether two texts were written by the same author or not (Koppel and Schler, 2004; Koppel et al., 2009).

The GLAD system of Hürlimann et al. (2015) was specifically developed to solve AV problems, and has been shown to be highly adaptable to new datasets (Halvani et al., 2018). GLAD uses an SVM with a variety of features including character level ones, which have proved to be most effective for AA tasks (Stamatatos, 2009; Moreau et al., 2015; Hürlimann et al., 2015), and is freely available. Moreover, Kestemont et al. (2019) show that many of the best models for authorship attribution are based on Support Vector Machines. Hence we adopt GLAD in the present study.

More specifically, we run GLAD on our datasets and study the interaction of four different dimensions: topic, gender, amount of evidence per author, and genre. In practice, we design intra-topic, cross-topic, and cross-genre experiments, controlling for gender and amount of evidence per author. The focus on cross-topic and cross-genre AV is in line with the PAN 2015 shared task (Stamatatos et al., 2015); this setting has been shown to be more challenging than the task definitions of previous editions (Juola and Stamatatos, 2013; Stamatatos et al., 2014).

[1] https://www.newyorker.com/culture/cultural-comment/the-unmasking-of-elena-ferrante

Contributions We advance AA for Italian introducing two novel datasets, *ForumFree* and *Diaries*, which contribute to enhance the amount of available Italian data suitable for AA tasks.[2]

Running a battery of experiments on personal writings, we show that AV is feasible even with little data, but more evidence helps. Gender and topic can be indicative clues, and if not controlled for, they might overtake more specific aspects of personal style.

2 Data

For the present study, we introduce two novel datasets, *ForumFree* and *Diaries*. Although already compiled (Maslennikova et al., 2019), the original ForumFree dataset was not meant for AA. Therefore, we reformat it following the PAN format[3]. The dataset contains web forum comments taken from the ForumFree platform[4], and the subset used in this work covers two topics, *Medicina Estetica* ("Aesthethic Medicine") and *Programmi Tv* ("Tv Programmes"; *Celebrities* in the original dataset). A third subset, *Mix*, is the union of the first two. The Diaries dataset is originally assembled for the present study, and contains a collection of diary fragments included in the project *Italiani all'estero: i diari raccontano* ("Italians abroad: the diaries narrate").[5] For Diaries, no topic classification has been taken into account. Table 1 shows an overview of the datasets.

Subset	# Authors			# Docs	W/A	D/A	W/D
	F	M	Tot				
Med Est	33	44	77	56198	63	661	48
Prog TV	78	71	149	153019	32	812	22
Mix	111	115	276	209217	41	791	29
Diaries	77	188	275	1422	462	5	477

Table 1: Overview of the datasets. W/A = Avg words per author; D/A = Avg docs per author; W/D = Avg words per doc.

2.1 Preprocessing

For the ForumFree dataset, comments which only contained the word *up*, commonly used on the internet to give new visibility to a post that was written in the past, were removed from the dataset, together with their authors when this was the only text associated with them.

The stories narrated in the diaries are of a very personal nature, which means that many proper nouns and names of locations are used. To avoid relying on these explicit clues, which are strong but not indicative of personal writing style, we perform Named Entity Recognition (NER), using *spaCy* (Honnibal, 2015). Person names, locations and organizations were replaced by their corresponding labels, namely *PER*, *LOC*, *ORG*. The fourth label used by *spaCy*, MISC (miscellany), was not considered; dates were also not normalized. Moreover, a separate set of experiments was performed by *bleaching* the diary texts prior to their input to the GLAD system. The bleaching method was proposed by van der Goot et al. (2018) in the context of cross-lingual Gender Prediction, and consists of transforming tokens into an abstract representation that masks lexical forms while maintaining key features. We only use 4 of the 6 original features. *Shape* transforms uppercase letters into 'U', lowercase ones into 'L', digits into 'D', and the rest into 'X'. *PunctA* replaces emojis with 'J', emoticons with 'E', punctuation with 'P' and one or more alphanumeric characters with a single 'W'. *Length* represents a word by the number of its characters. *Frequency* corresponds to the *log* frequency of a token in the dataset. The features are then concatenated. The word 'House' would be rewritten as 'ULLLL W 05 6'.

2.2 Reformatting

We reformat both datasets in order to make them suitable for AV. The data is divided into so-called *problems*: each problem is made of a known and an unknown text of equal length.

To account for the shortness of the texts and to avoid topic biases that would derive by taking consecutive text as known and unknown fragments, all the documents written by the same author are first shuffled and then concatenated into a single string. The string is split into two spans containing the same number of words, so that the words contained in the unknown span come from subsets of texts which are different from the ones that form the known one. An example of this process is displayed in Figure 1. Rather than being represented by individual productions, each author is therefore represented by a *set* of texts, whose original se-

[2]Further information about the datasets can be found at https://github.com/garuggiero/Italian-Datasets-for-AV

[3]https://pan.webis.de/clef15/pan15-web/authorship-verification.html

[4]https://www.forumfree.it/

[5]https://www.idiariraccontano.org

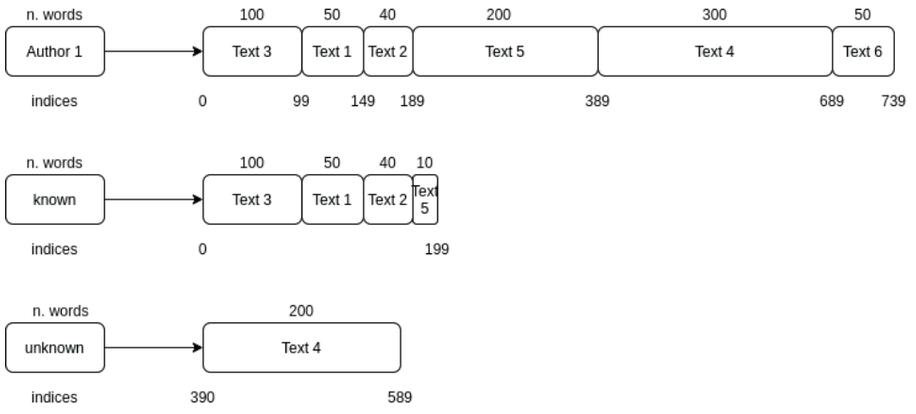

Figure 1: Example of the creation of known and unknown documents for the same author when considering 400 words per author.

quential order has been altered. Each known text is paired with an unknown text from the same author. To create negative instances, given a dataset with multiple problems, one can (i) make use of external documents (*extrinsic* approach (Seidman, 2013; Koppel and Winter, 2014)), or (ii) use fragments collated from all authors in the training data, except the target author (*intrinsic* approach). We create negative instances with an intrinsic approach. More specifically, following Dwyer (2017), the second half of the unknown array is shifted by one, so that the texts of the second half of the known array are paired with a *different-author* text in the unknown array. In this way, the label distribution is balanced.

3 Method

Given a pair of known and unknown fragments (KU pair), the task is to predict whether they are written by the same author or not. In designing our experiments, we control for topic, gender, amount of evidence, and genre. The latter is fostered by the diverse nature of our datasets.

Topic Maintaining the topic roughly constant should allow stylistic features to gain more discriminative value. We design intra-topic (IT) and cross-topic experiments (CT). In IT, we distinguish same- and different-topic KU pairs. In same-topic, we train and test the system on KU pairs from the same topic. In different-topic, we include the Mix set and the diaries. Since we train and test on a mixture of topics and there can be topic overlap, these are not truly cross-topic, and we do not consider them as such.

Given that no topic classification is available for the diaries, the CT experiments are only performed on the ForumFree dataset. We train the system on Medicina Estetica and test it on Programmi Tv, and vice versa.

Gender Previous work has shown that similarity can be observed in writings of people of the same gender (Basile et al., 2017; Rangel et al., 2017).[6] In order to assess the influence of same vs different gender in AA, we consider three gender settings: only female authors and only male authors (*single-gender*), and *mixed-gender*, where the known and unknown document can be either written by two authors of the same gender, or by a male and a female author. In dividing the subsets according to the gender of the authors, we consider gender implicitly. However, we also perform experiments adding gender as feature to the instance vectors, indicating both the gender of the known and unknown documents' authors and whether or not the gender of the authors is the same.

Evidence Following Feiguina and Hirst (2007), we experiment with KU pairs of different sizes, i.e. with 400, 1 000, 2 000 and 3 000 words per author. Each element of the KU pair is thus made up of 200, 500, 1 000 and 1 500 words respectively. To observe the effect of the different text sizes on the classification, we manipulate the number of instances in training and test, so that the same authors are included in all the different word settings of a single topic-gender experiment.

[6]Binary gender is a simplification of a much more nuanced situation in reality. Following previous work, we adopt it for convenience.

Genre We perform cross-genre experiments (CG) by training on ForumFree and testing on the Diaries, and vice versa.

Splits and Evaluation We train on 70% and test on 30% of the instances. However, since we are controlling for gender and topic, the number of instances contained in the training and test sets varies in each experiment. We keep the test sets stable across IT, CT and CG experiments, so that we can compare results. Following the PAN evaluation settings (Stamatatos et al., 2015), we use three metrics. $c@1$ takes into account the number of problems left unanswered and rewards the system when it classifies a problem as unanswered rather than misclassifying it.

Probability scores are converted to binary answers: every score greater than 0.5 becomes a positive answer, every score smaller than 0.5 corresponds to a negative answer and every score which is exactly 0.5 is considered as an unanswered problem. The AUC measure corresponds to the area under the ROC curve (Fawcett, 2006), and tests the ability of the system to rank scores properly, assigning low values to negative problems and high values to positive ones (Stamatatos et al., 2015). The third measure is the product of $c@1$ and AUC.

Model We run all experiments using GLAD (Hürlimann et al., 2015). This is an SVM with *rbf* kernel, implemented using Python's *scikit-learn* (Pedregosa et al., 2011) library and NLTK (Bird et al., 2009). GLAD was designed to work with 24 different features, which take into account stylometry, entropy and data compression measures. We compare GLAD to a simple baseline which randomly assigns a label from the set of possible labels (i.e. 'YES' or 'NO') to each test instance.

Our choice fell on GLAD for a variety of reasons. As a general observation, even in later challenges, SVMs have proven to be the most effective for AA tasks (Kestemont et al., 2019). More specifically, in a survey of freely available AA systems, GLAD showed best performance and especially high adaptability to new datasets (Halvani et al., 2018). Lastly, de Vries (2020) has explored fine-tuning a pre-trained model for AV in Dutch, a less-resourced language compared to English. He found that fine-tuning BERTje (a Dutch monolingual BERT-model, (de Vries et al., 2019)) with PAN 2015 AV data (Stamatatos et al., 2015), failed to outperform a majority baseline (de Vries, 2020). He concluded that Tranformer-encoder models might not suitable for AA tasks, since they will likely overfit if the documents contain no reliable clues of authorship (de Vries, 2020).

4 Results and Discussion

The number of experiments is high due to the interaction of the dimensions we consider.

Tables 2 and 3 only include the mixed-gender results of the IT experiments on Mix (which corresponds to the entire ForumFree dataset used for this study) and Diaries, respectively. Results concerning all dimensions considered are anyway discussed in the text. We refer to the combined score. Since the baseline results are different for each setting, we do not include them. However, all models perform consistently above their corresponding baseline.

For the Mix topic, we achieved 0.966 with 96 authors in total and 3 000 words (Table 2). For the diaries, we achieved 0.821 with 46 authors in total and 3 000 words each (Table 3).[7] Although the training and test sets are of different sizes for both datasets, more evidence seems to help the model to solve the problem.

In the IT experiments, the highest score for Medicina Estetica is 0.923, with 41 authors in total and 1 000 words per author, and for Programmi Tv 0.944, with 59 authors and 3 000 words each. In the CT setting, the scores stay basically the same in both directions. In CG, when training on the diaries and testing on Mix, we obtain the same score when training on Mix with 3 000 words. When training on Mix and testing on Diaries, we achieved 0.737 on the same test set, and 0.748 with 1 000 words per instance.

Discussion When more variables interact in the same subset, as in mixed-gender sets of the ForumFree and Diaries dataset, we found that the classifier uses the implicit gender information. Indeed, it achieves slightly better scores in mixed-gender settings than in female- and male-only ones, suggesting that the classifier might be using internal clustering of the data rather than writing style characteristics. This also explains why results are higher in Mix than in separate topics, because the classifier can use topic information.

[7] Using a bleached representation of the texts, the score increased by 0.36

# W/A	# Auth	# Problems		Eval					
		Train	Test	C	I	U	c@1	AUC	*
400	127	88	39	33	6	0	0.846	0.947	0.801
1 000	109	76	33	30	3	0	0.909	0.926	0.842
2 000	100	70	30	29	1	0	**0.967**	0.995	0.962
3 000	96	67	29	28	1	0	0.966	**1.000**	**0.966**

Table 2: Training and test set configurations and IT evaluation scores on Mix texts written by female and male authors. *C,I* and *U* are Correct, Incorrect, Unanswered problems.

# W/A	# Auth	# Problems		Eval					
		Train	Test	C	I	U	c@1	AUC	*
400	229	160	69	47	21	1	0.691	0.725	0.500
1 000	180	126	54	43	11	0	0.796	0.891	0.709
2 000	98	68	30	25	5	0	**0.833**	0.905	0.754
3 000	46	32	14	12	2	0	0.857	**0.958**	**0.821**

Table 3: Training and test configurations and IT evaluation scores on diaries made of NE converted text written by both genders. *C,I* and *U* are Correct, Incorrect, Unanswered problems.

We also observe that by adding gender as an explicit feature in topic- and gender-controlled subsets, GLAD uses this information to improve classification, especially in mixed-gender scenarios.

Although previous research demonstrated that CT and CG experiments are harder than IT ones (Sapkota et al., 2014; Stamatatos et al., 2015), in our case the scores for the three settings are comparable. However, since we only performed CT and CG experiments on mixed-gender subsets, the gender-specific information might have also played a role in this process (see above).

Overall, the experiments show that using a higher number of words per author is preferable. Although 3 000 words seems to be optimal for most settings, in the large number of experiments that we carried out (not all included in this paper) we also observed that lower amounts of words also led to comparable results. This aspect will require further investigation.

5 Conclusion

We experimented with AV on Italian forum comments and diary fragments. We compiled two datasets and performed experiments which considered the interaction among topic, gender, length and genre. Even when the texts are short and present more individual variation than traditional texts used in AA, AV is a feasible task, but having more evidence per author improves classification. While making the task more challenging, controlling for gender and topic ensures that the system prioritizes authorship over different data clusters. Although the datasets used are intended for AV problems, they can be easily adapted to other AA tasks. We believe this to be one of the major contributions of our work, as it can help to advance the up-to-now limited AA research in Italian.

Acknowledgments

The ForumFree dataset was a courtesy of the Italian Institute of Computational Linguistics "Antonio Zampolli" (ILC) of Pisa.[8]

References

Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. 2017. N-GrAM: New Groningen Author-profiling Model—Notebook for PAN at CLEF 2017. In *CEUR Workshop Proceedings*, volume 1866.

Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Martin Potthast, Francisco Rangel, Paolo Rosso, Günther Specht, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, and Eva Zangerle. 2020. Shared Tasks on Authorship Analysis at PAN 2020. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*,

[8] http://www.ilc.cnr.it/

pages 508–516, Cham. Springer International Publishing.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.

Wietse de Vries. 2020. Language Models are not just English Anymore: Training and Evaluation of a Dutch BERT-based Language Model Named BERTje. Master Thesis in Information Science, University of Groningen, The Netherlands.

Gareth Terence Bryan Dwyer. 2017. Novel approaches to authorship attribution. Master Thesis in Language and Communication Technologies, Information Science, University of Groningen, The Netherlands.

Tom Fawcett. 2006. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Olga Feiguina and Graeme Hirst. 2007. Authorship attribution for small texts: Literary and forensic experiments. In *Proceedings of the SIGIR'07 Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 2007)*.

Oren Halvani, Christian Winter, and Lukas Graner. 2018. Unary and binary classification approaches and their implications for authorship verification. *arXiv preprint arXiv:1901.00399*.

Matthew Honnibal. 2015. spacy: Industrial-strength natural language processing (nlp) with python and cython.

Manuela Hürlimann, Benno Weck, Esther van den Berg, Simon Suster, and Malvina Nissim. 2015. Glad: Groningen lightweight authorship detection. In *CLEF (Working Notes)*.

Patrick Juola and Efstathios Stamatatos. 2013. Overview of the Author Identification Task at PAN 2013. *CLEF (Working Notes)*, 1179.

Patrick Juola. 2015. The Rowling case: A proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities*, 30(suppl_1):i100–i113.

Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. 2019. Overview of the Cross-domain Authorship Attribution Task at PAN 2019. In *CLEF (Working Notes)*.

Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. Overview of the Cross-Domain Authorship Verification Task at PAN 2020. In Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol, editors, *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org, September.

Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*, page 62.

Moshe Koppel and Yaron Winter. 2014. Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1):178–187.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26.

Aleksandra Maslennikova, Paolo Labruna, Andrea Cimino, and Felice Dell'Orletta. 2019. Quanti anni hai? Age Identification for Italian. In *Proceedings of 6th Italian Conference on Computational Linguistics (CLiC-it), 13-15 November, 2019, Bari, Italy.*

Erwan Moreau, Arun Jayapal, Gerard Lynch, and Carl Vogel. 2015. Author verification: basic stacked generalization applied to predictions from a set of heterogeneous learners-notebook for pan at clef 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org.

Abdulfattah Omar, Basheer Ibrahim Elghayesh, and Mohamed Ali Mohamed Kassem. 2019. Authorship attribution revisited: The problem of flash fiction a morphological-based linguistic stylometry approach. *Arab World English Journal (AWEJ) Volume*, 10.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working notes papers of the CLEF*, pages 1613–0073.

Upendra Sapkota, Thamar Solorio, Manuel Montes, Steven Bethard, and Paolo Rosso. 2014. Cross-topic authorship attribution: Will out-of-topic data help? In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1228–1237.

Shachar Seidman. 2013. Authorship verification using the impostors method. In *CLEF 2013 Evaluation labs and workshop–Working notes papers*, pages 23–26. Citeseer.

Efstathios Stamatatos and Moshe Koppel. 2011. Plagiarism and authorship analysis: introduction to the special issue. *Language Resources and Evaluation*, 45(1):1–4.

Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A Sanchez-Perez, and Alberto Barrón-Cedeño. 2014. Overview of the author identification task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014*, pages 1–21.

Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Patrick Juola, Aurelio López-López, Martin Potthast, and Benno Stein. 2015. Overview of the author identification task at pan 2015. clef 2015 evaluation labs and workshop, online working notes, toulouse, france. In *CEUR Workshop Proceedings*, pages 1–17.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Michail Tsikerdekis and Sherali Zeadally. 2014. Multiple account identity deception detection in social media using nonverbal behavior. *IEEE Transactions on Information Forensics and Security*, 9(8):1311–1321.

Arjuna Tuzzi and Michele A Cortelazzo. 2018. *Drawing Elena Ferrante's Profile: Workshop Proceedings, Padova, 7 September 2017*. Padova UP.

Rob van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction. *arXiv preprint arXiv:1805.03122*.

Min Yang and Kam-Pui Chow. 2014. Authorship attribution for forensic investigation with thousands of authors. In *IFIP International Information Security Conference*, pages 339–350. Springer.